Chiem MAT 115 Homework 4

As per usual, for this assignment you should write your answers in RMarkdown but submit both Markdown and knit pdf files. The due date for this assignment is **Oct. 9 at 2:30 PM**.

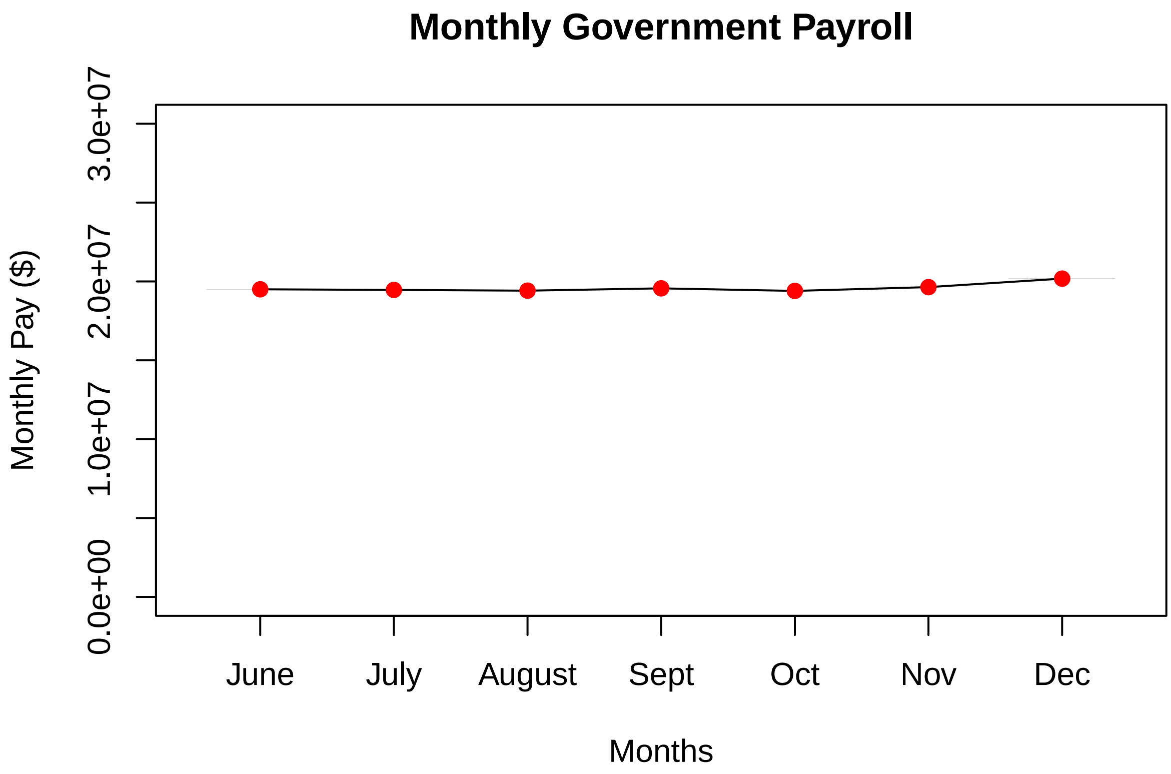
Part 1

For part 1 of this assignment you will recreate a government payroll plot I showed you on the first day of class. It will also be part of a future reading. Here is a link to the reading: "How To Lie With Statistics". The plot is on page 65.

Unlike previous assignments, this dataset is not available in an R package. I have made it available to you as a Google Sheet. You can upload data from a Sheet using the `googlesheets4` package (`read_sheet` function) and the link to the sheet. Note, you may have to go through some authorization steps to access the Sheet.

```r
require(googlesheets4)
library(tidyverse)
gs4_auth(email = "dchiem@fandm.edu") # uses my school email for authorization

link <- paste("https://docs.google.com/spreadsheets/d/1W_FBpViIChZB-Yv9m1q3qrM85A",
              "4seNL7zjDjGa7zoaA/edit?usp=sharing",sep="")

govpay <- read_sheet(link)
```

```r
months_order <- unique(govpay$month) # gets the month in order

# used to find the sum of the payrolls for each month.
# arranges it using the months_order vector
govpay_bymonth <- govpay %>%
  group_by(month) %>%
  summarize(Monthly_Pay = sum(payroll), .groups='drop') %>%
  arrange(match(month, months_order))

months <- factor(govpay_bymonth$month, levels = months_order) # makes it a factor
#so I can use it to create a plot

# graph with the scale as min value to max value.
# Makes it look like it has a large growth
#plot(months, govpay_bymonth$Monthly_Pay, type='n',
     #main = "Monthly Government Payroll",
     #xlab = "Months", ylab = "Monthly Pay ($)", lty=0)

#lines(months, govpay_bymonth$Monthly_Pay)
#points(months, govpay_bymonth$Monthly_Pay, col ="lightblue", pch=19)

# plots the months by the monthly pay. The y scale here is from 0 to 30,000,000
plot(months, govpay_bymonth$Monthly_Pay, type='n',
     main = "Monthly Government Payroll",
     xlab = "Months", ylab = "Monthly Pay ($)", lty=0,
     ylim=c(0,30000000))

# draws lines and points separatley to ensure proper formatting.
# Without this each point would be a solid line.
lines(months, govpay_bymonth$Monthly_Pay)
points(months, govpay_bymonth$Monthly_Pay, col ="red",
       pch=19)
```



The dataset comes in the form of daily pay. It is your responsibility to wrangle the data so that it is on the monthly x-axis scale.

You also need to determine the appropriate y-axis scale. And justify it. How could you mislead your audience by changing the y-axis scale?

A: I can mislead them by changing the y limit. If I make the scale smaller (First graph), superficially it looks like the pay goes up by a larger amount. If I use a larger scale (Second graph), it makes it look like the pays have stayed relatively the same.

In this case, I used the larger scale because I think it more accurately describes the growth. The data stays between ~19.3 million and ~20.2 million which is not as exponentially large of a growth as the graph with the smaller scale makes it seem.

What additional information could help to justify one y-axis scale or another?

A: Having past payrolls can be useful in determining the actual rate of how salaries are increasing or decreasing. Having that data would make it easier to adjust the scale to accurately reflect the historical rate.

Be sure to include all wrangling AND plotting code.

Optional (aka not graded): Figure out how to add an image to the background of the plot similar to the national income plots from the reading (start on pg. 61).

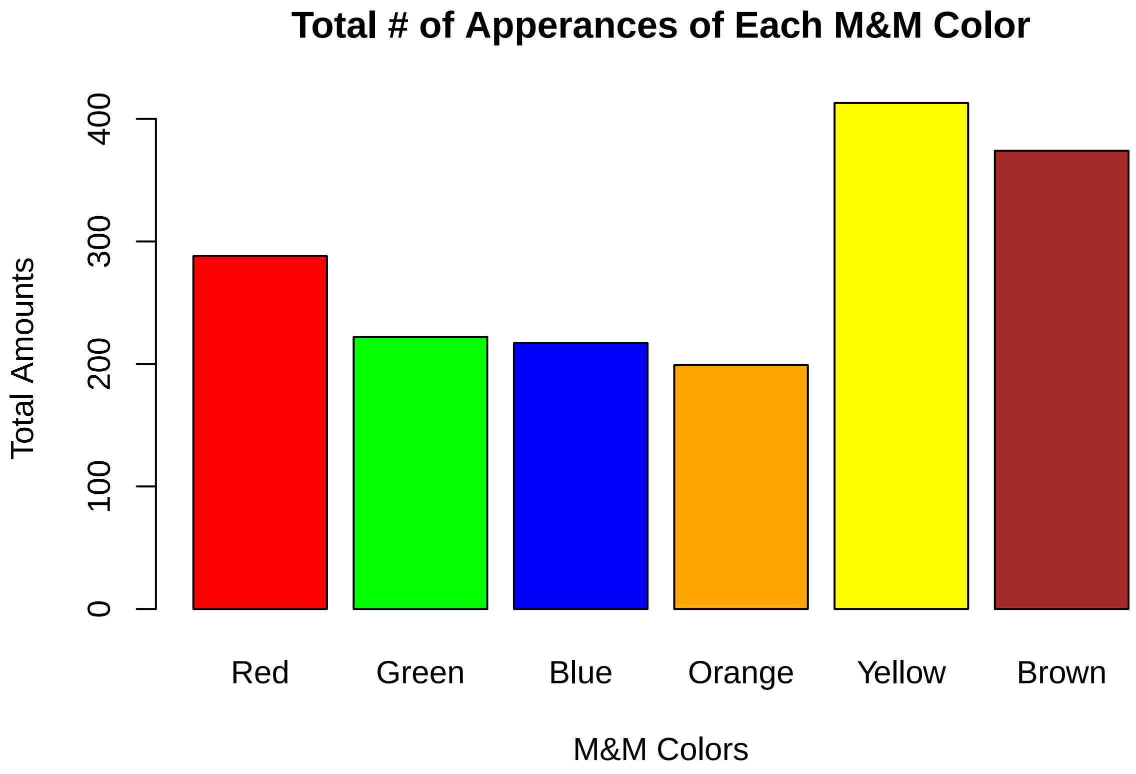
Part 2

For part 2, you are the resident data scientist of the Mars, Inc. candy company. They are asking you to provide a quality control report for one of the companies most popular candies - M&M's.

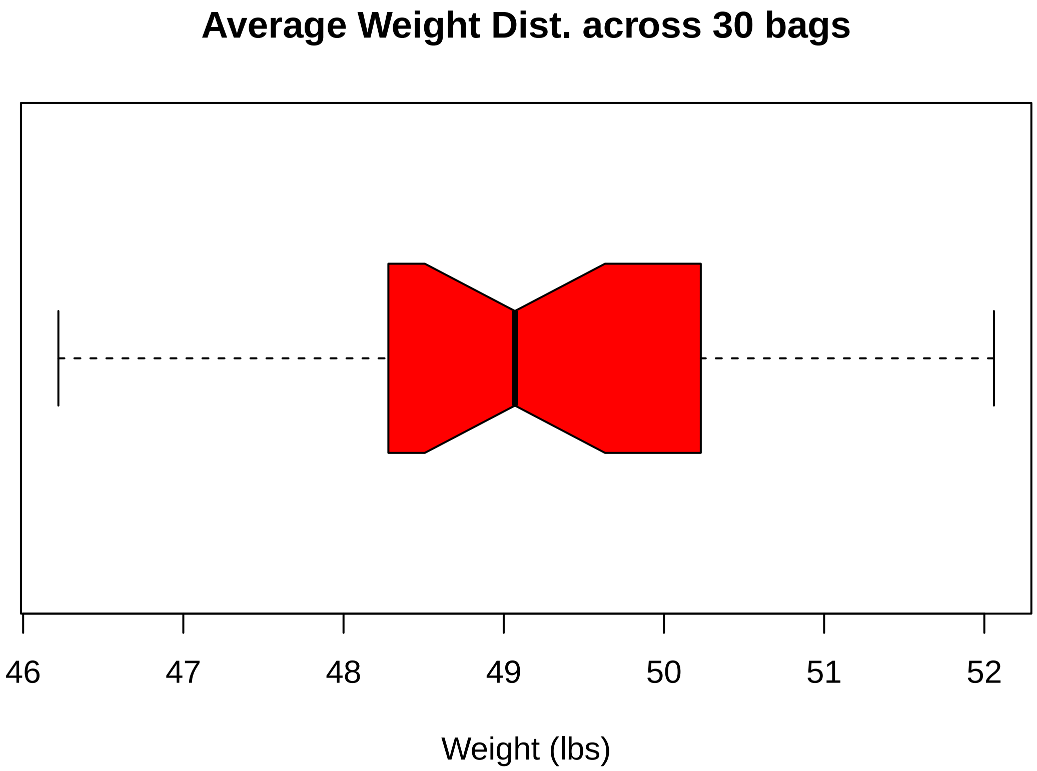
The powers that be are worried the candies are not getting distributed evenly based on color *and* the number of candies per bag. They want you to determine if their worries are founded. Your co-workers have already collected the necessary data by tallying the frequencies of each M&M color for 30 bags and organizing the data in a Google Sheet.

```
mandm <- read_sheet(paste("https://docs.google.com/spreadsheets/d/1lIkKV3QbhLryb8",
                          "5TvB78DNv8w2HMO4KP6hklsz5Ayio/edit?usp=sharing",sep=""))
```

```
#Total apperances of all colors

color_sums <- vector(length = (ncol(mandm) - 1)) # holds the sum of each color appearance
color_names <- vector(length = (ncol(mandm) - 1)) # holds the name of each color

# for loop is used to iterate over the columns
for (i in 1:length(color_sums)) {
  color_names[i] <- colnames(mandm)[i]
  color_sums[i] <- sum(mandm[[i]])
}

# used for color of each bar
bar_colors = c("red", "green", "blue", "orange", "yellow", "brown")

# barplot that shows the relationship between color
# and the frequency of that color in each bag
barplot(color_sums,
        names.arg = color_names,
        main = "Total # of Apperances of Each M&M Color",
        col = bar_colors,
        xlab = "M&M Colors",
        ylab = "Total Amounts")
```



```r
#Weight distribution in order to figure out # of candies per bag is consistent

# box and whisker plot to show weight distribution.
# There is only one variable in this case which is why I used a boxplot.
boxplot(mandm$Weight,
        main = "Average Weight Dist. across 30 bags",
        horizontal=TRUE,
        xlab = "Weight (lbs)",
        notch = TRUE,
        col = "red")
```



The primary way you will confirm or assuage your employer's fears are with visualizations. So, you must create AT LEAST two plots and written explanations for any patterns (or lack thereof) within the plots as they pertain to your employer's fears. Remember, your bosses are NOT data scientists, statisticians, or computer scientists. Be sure that your plots and explanations are simple, clear, and convincing!

A: (In a total of 30) There appears to be an unequal amount of yellow and brown when compared to blue, green, orange, and red. The color that appears the most is yellow with it (and brown) appearing far more frequently than the other 4 colors. According to the bargraph there is an unequal distribution of colors.

A2 (For weight distribution): The box and whisker plot demonstrates that the weight of each bag is slightly skewed towards Q1, meaning that the weight distributions tend to be a bit lighter. There are no outliers indicating that there are no unusual distributions. In addition, the box itself is wide, which suggests that the weights tend to vary. This would mean that the number of m&ms per bag are inconsisten.

Be sure to include all wrangling AND plotting code.

For this part, include a step-by-step written annotation of your data wrangling code. What does each function and/or line of code do?

A: For the visualization of color frequencies, I first created two vectors. One was to store the total apperances of each color (in order of the columns), and the color names. I then used a for loop to iterate over each column and stored the name and sum in their respective vectors. I used a bar graph to visualize the frequencies. The y axis represents the apperances, and the x axis represents each color. I changed the bar colors to match their corresponding colors in order to make it easier to look at. I this by setting col to a bar_colors vector that I created that contains each color

For the weight distribution visualization, I used a box and whisker plot. I used mandm$weight for the first argument which will be on the x-axis in this case since I made the box plot horizontal. THe x-axis represents the weight in lbs of each bag. I added a notch to the box because it makes it easier to see where the median is skewed towards. I made the box red in order to emphasize it more.

What are the ethical considerations for this data and analysis?

A: I think this could lead to more exploitation of the consumer. The company could use this data to make the bags more consistently large, and advertise it that way in order to make consumers purchase this. Since m&ms use a lot of synthesized sugars and fats (Which allow it to be more addicting as well), this could also negatively impact consumer health.

If production is greatly increased due to the company's goal of providing larger bags, it could also negativley impact the environment.

Optional (aka not graded): What statistical test or tests could help with interpretation of these data?